
BIG
DATA
ALSO
TOOLS
DISK
TARGET
EXAMPLES
EVERY
APPLIED
SENSOR
SHARED
DEFINITION
CURRENT
MAY
MOVING
WIRELESS
MANAGE
ARCHIVES
WITHIN
BIOGEOCHEMICAL
TECHNOLOGIES
USING
CONNECTOMICS
ORGANIZATIONS
SET
USE
RELATIONAL
LARGER
CONTINUES
SOCIAL
INDEXING
CITATION
TENS
COMPLEX
ANALYTICS
RECONSIDER
ZETTABYTES
PRACTITIONERS
CAPTURE
THOUGHT
INTERNET
DISTRIBUTED
BUSINESS
SETS
GENOMICS
DESCRIBING
RADIO-FREQUENCY
MANAGEMENT
TERABYTES
CASE
COMPLEXITY
NEEDED
REQUIRING
TIME
INFORMATION
ONE
LARGE
PROCESS
USED
CREATED
NOW
BIOLOGICAL
PROCESSING
UBIQUITOUS
CAPACITY
BURIED
WORLD'S
HUNDREDS
DESKTOP
NETWORKS
DATABASES
RECORDS
SOLID
TYPES
GARTNER
CURRENTLY
WORKING
AMOUNT
OPPORTUNITIES
ELAPSED
FORMS
PETABYTES
SYSTEMS
INCLUDE
TOLERABLE
ABILITY
SOFTWARE
PERFORMANCE
STORE
PRESENTATIONS
COMBAT
SIZE
RESEARCH
STORAGE
MPP
LOGS
COST
SIGNIFICANT
DEFINING
RELATED
COMPUTING
SAN
QUALITIES
PARALLEL
SINCE
MASSIVELY
GROW
SEARCH
DIFFICULTY

WHAT IS BIGDATA?

Data sets that exceed the boundaries and sizes of normal processing capabilities, forcing you to take a non-traditional approach

Big Data characteristics

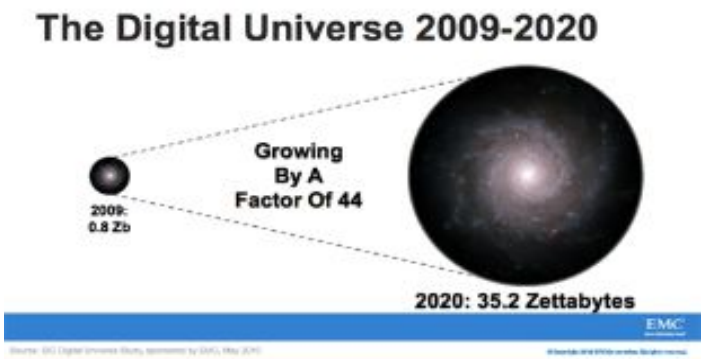
Volume: Describes the amount of data generated by organizations or individuals.

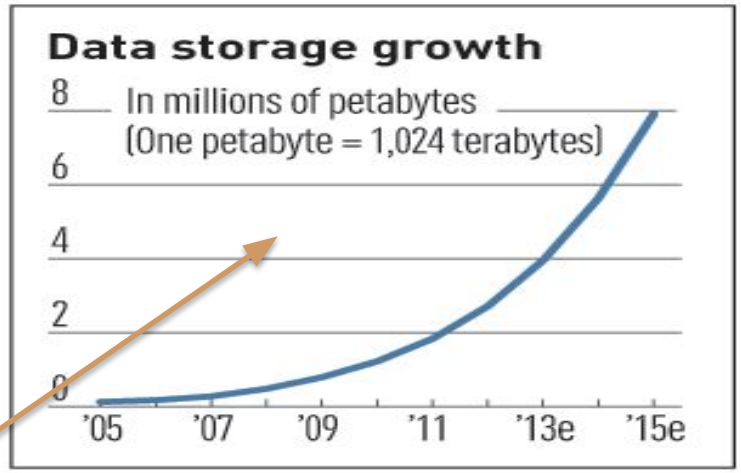
Variety: Describes structured and unstructured data.

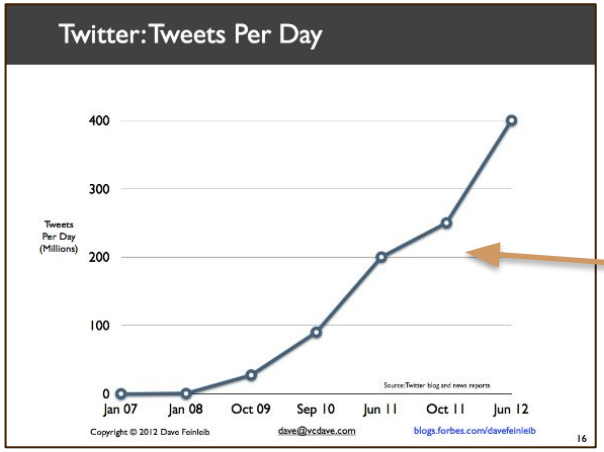
Velocity: Describes the frequency at which data is generated, captured and shared.

Volume (Scale)

- ✓ **Data Volume**
  - 44x increase from 2009 to 2020
  - From 8 zettabytes to 35 zb
- ✓ Data volume is increasing exponentially







Exponential increase in collected/generated data

Model of Generating/Consuming Data

Old Model: Few companies are generating data, all others are consuming data

New Model: all of us are generating data, and all of us are consuming data

Who is generating Big Data?

Social

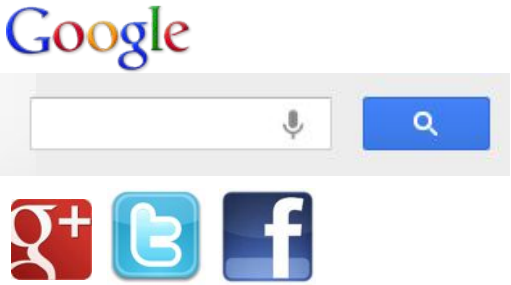
User Tracking & Engagement

Homeland Security

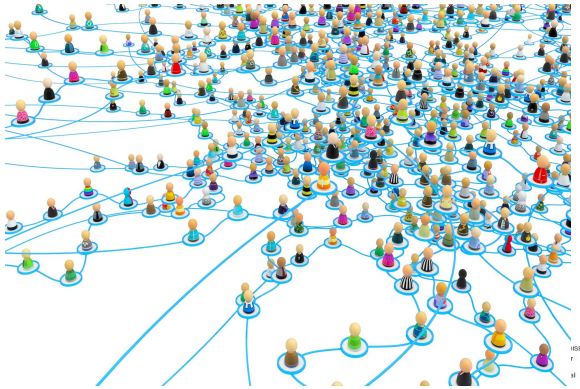
eCommerce

Financial Services

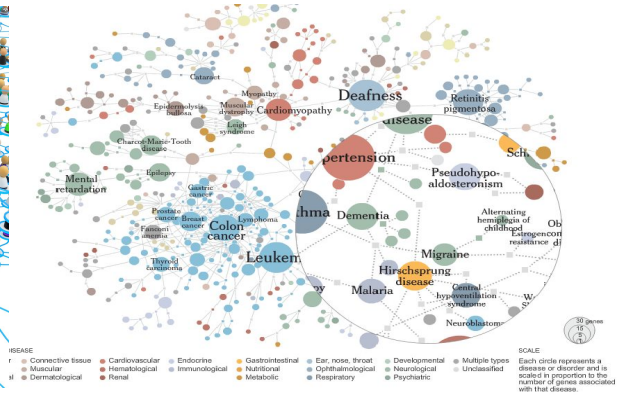
Real Time Search



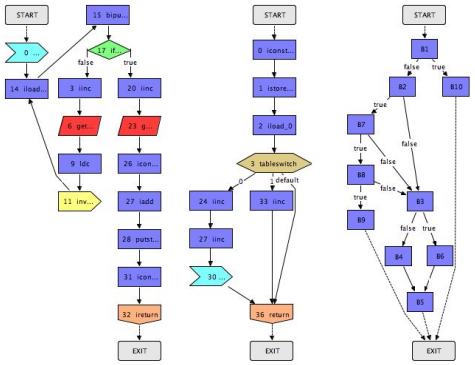
… and no data is an island



social networks

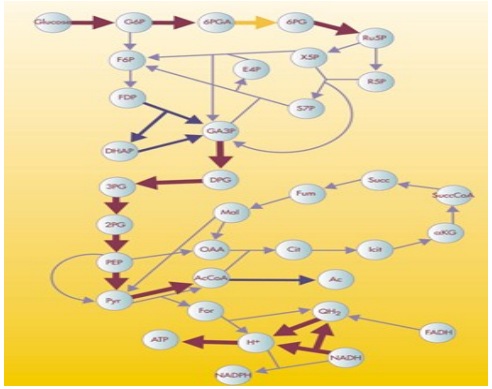


knowledge graph

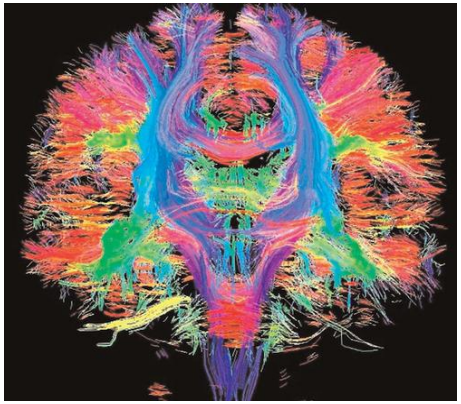


control flow graph

cyber networks



metabolic networks



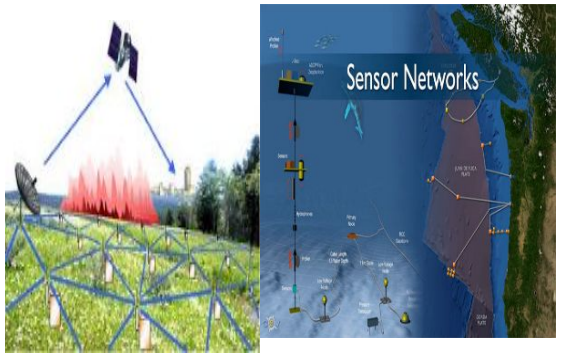
brain network

Big Data sources

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)



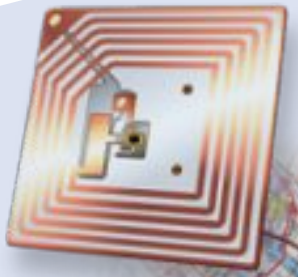
Sensor technology and networks
(measuring all kinds of data)

18+ TBs
of tweet data every day

? TBs of data every day

35+ TBs **of log data every day**

60 billion RFID tags today



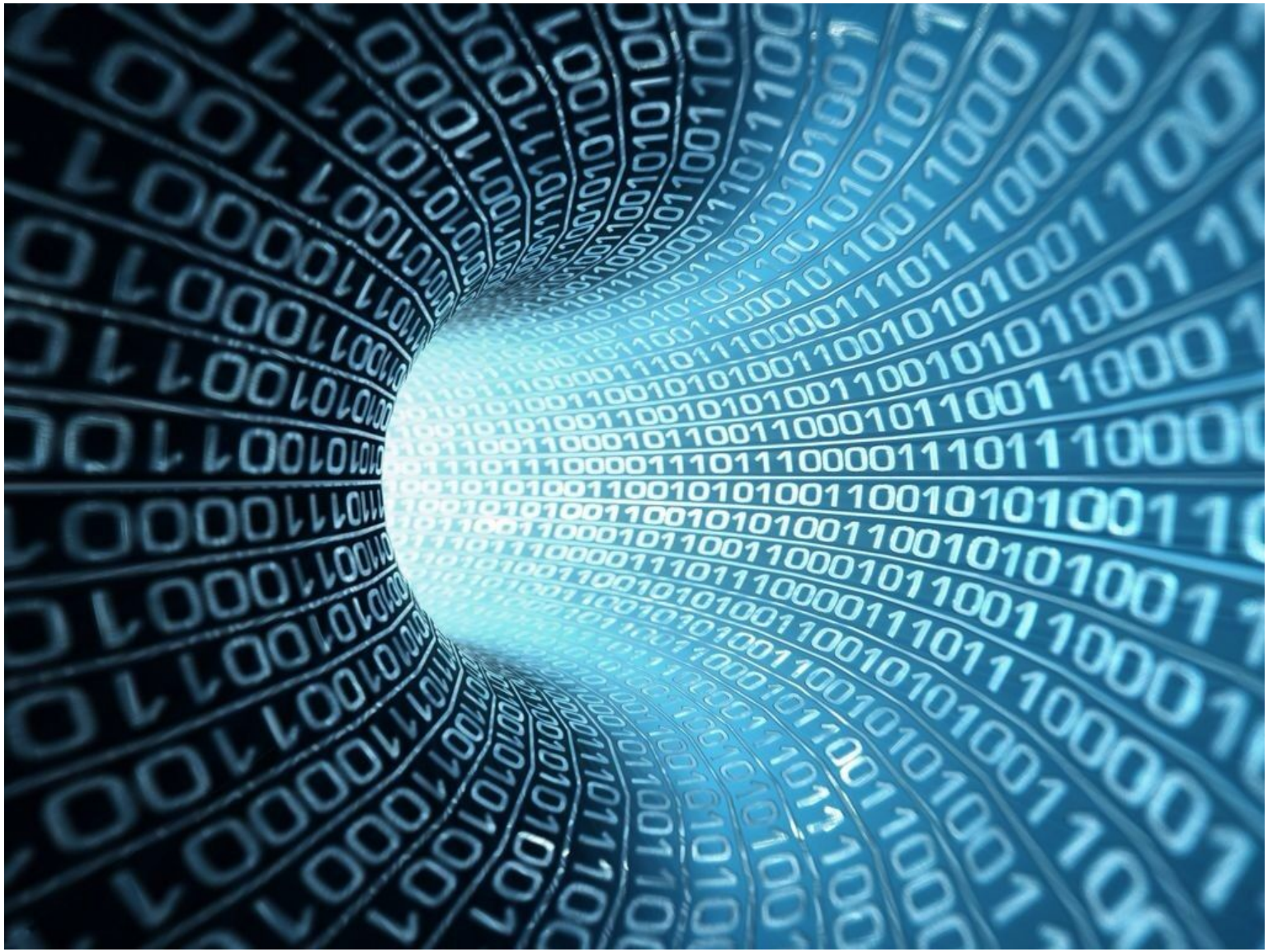
4.6 billion camera phones world wide

100s of millions of GPS enabled devices sold annually

76 million smart meters

10+ billion people on the Web

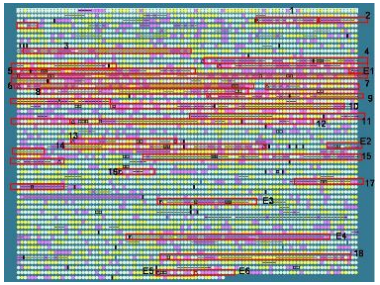
That is a lot of data …

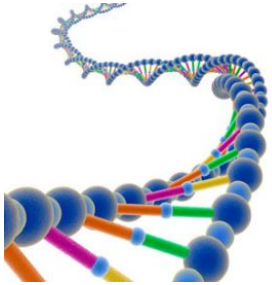


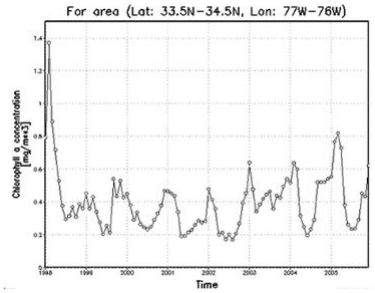
Variety (Complexity)

n Different Types:

- ✓ Relational Data (Tables/Transaction/Legacy Data)
- ✓ Text Data (Web)
- ✓ Semi-structured Data (XML)
- ✓ Graph Data
  - – Social Network, Semantic Web (RDF), …

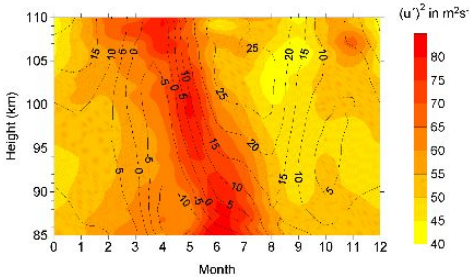






- ✓ Streaming Data
  - – You can only scan the data once

- ✓ A single application can be generating/collecting many types of data

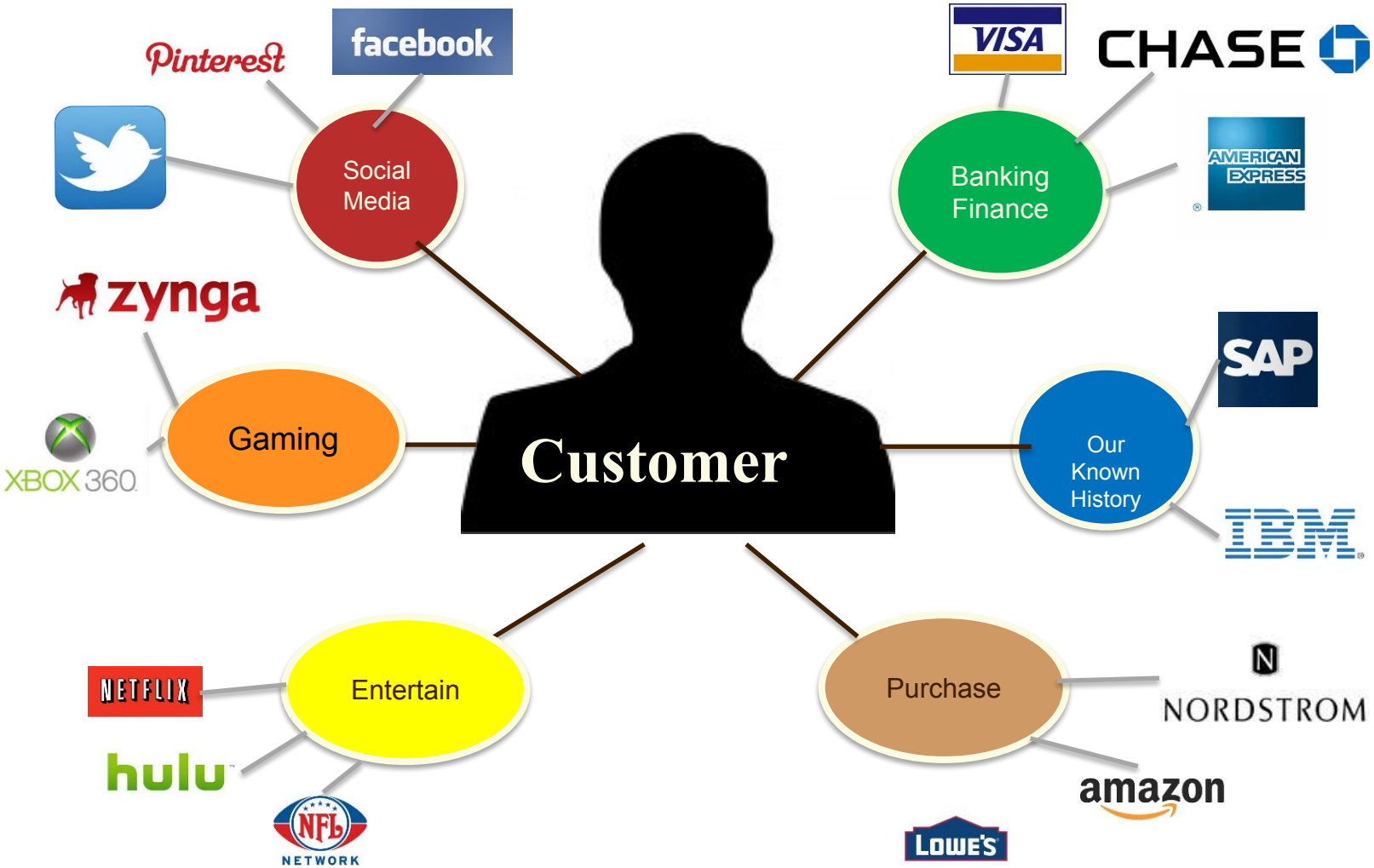


To extract knowledge→ all these types of data need to linked together

Structured data vs Unstructured data

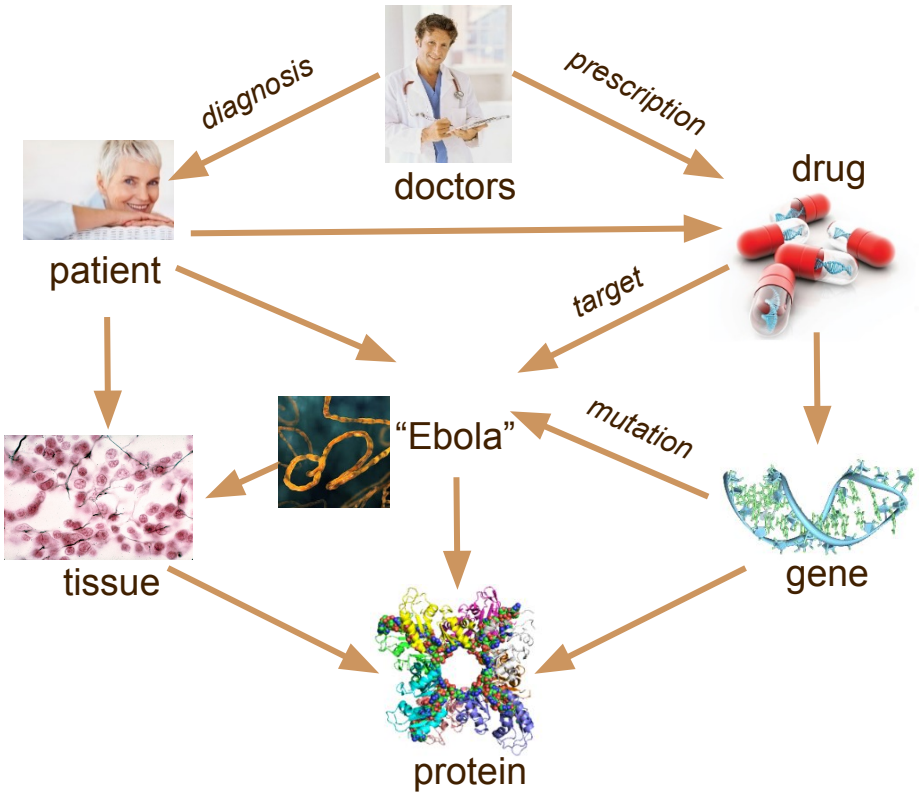
Traditional approach is no more sufficient to handle today's big data

A Single View to the Customer



A Global View of Linked Big Data

Diversified social network



Heterogeneous information network

Velocity (Speed)

- ✓ Data is begin generated fast and need to be processed fast
- ✓ Online Data Analytics
- ✓ Late decisions → missing opportunities

Google

Processes 20 PB a day (2008)
Crawls 20B web pages a day (2012)
Search index is 100+ PB (5/2014)
Bigtable serves 2+ EB, 600M QPS (5/2014)

JPMorganChase

150 PB on 50k+ servers running 15k apps (6/2011)

YAHOO!

Hadoop: 365 PB, 330K nodes (6/2014)

amazon web services™

S3: 2T objects, 1.1M request/second (4/2013)

ebay

Hadoop: 10K nodes, 150K cores, 150 PB (4/2014)

300 PB data in Hive + 600 TB/day (4/2014)

facebook

INTERNET ARCHIVE

400B pages, 10+ PB (2/2014)

How much data?

Data never sleeps…

How Much Data Is Generated Every Minute? 24/7/365

Email Users Send 20,41,66,667 Emails

Data never sleeps…

How Much Data Is Generated Every Minute? 24/7/365

Google Receives Over 20,00,000 Search Queries

Data never sleeps…

How Much Data Is Generated Every Minute? 24/7/365

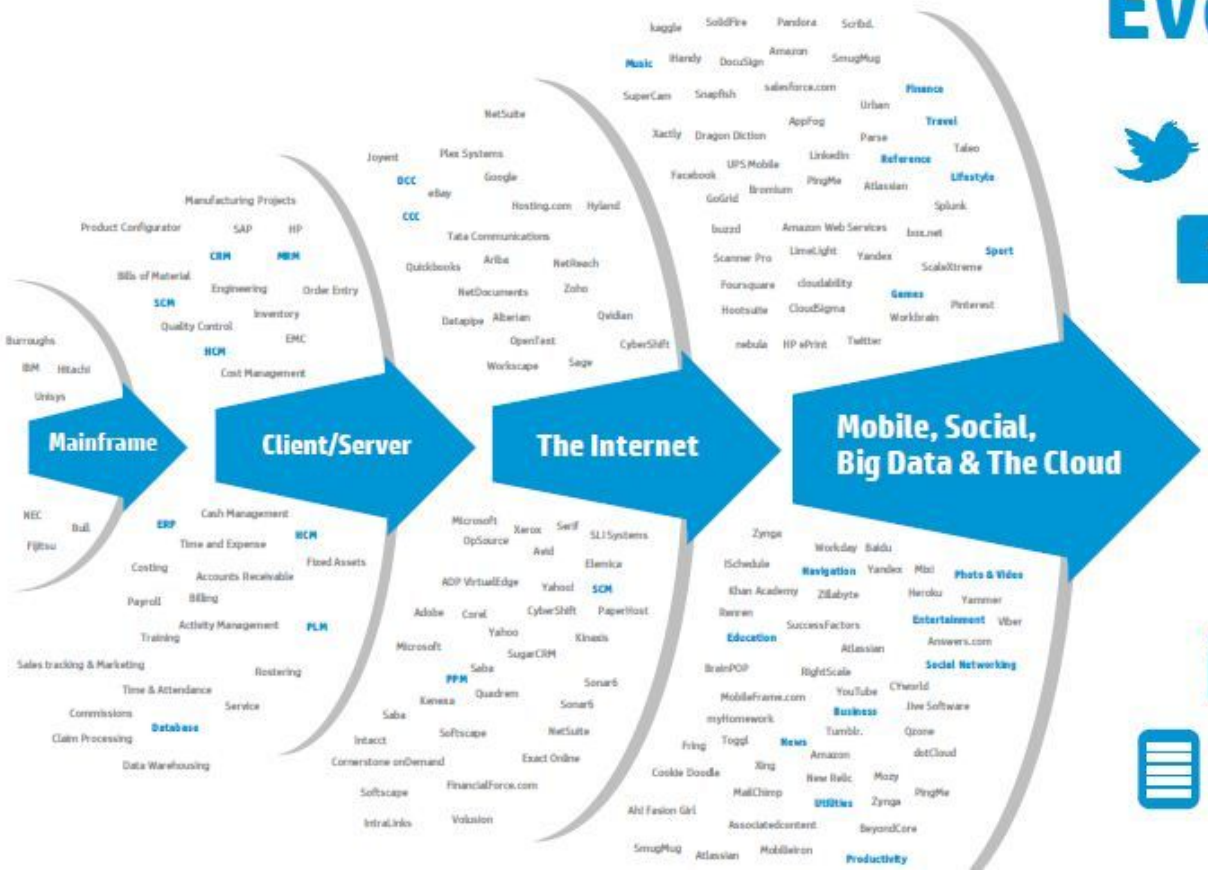
Apple Receives About 47,000 App Downloads

Data never sleeps…

How Much Data Is Generated Every Minute? 24/7/365

Brands on Facebook Get 34,722 Likes

A new style of IT emerging



Every 60 seconds

98,000+ tweets

695,000 status updates

11million instant messages

698,445 Google searches

168 million+ emails sent

1,820TB of data created

217 new mobile web users

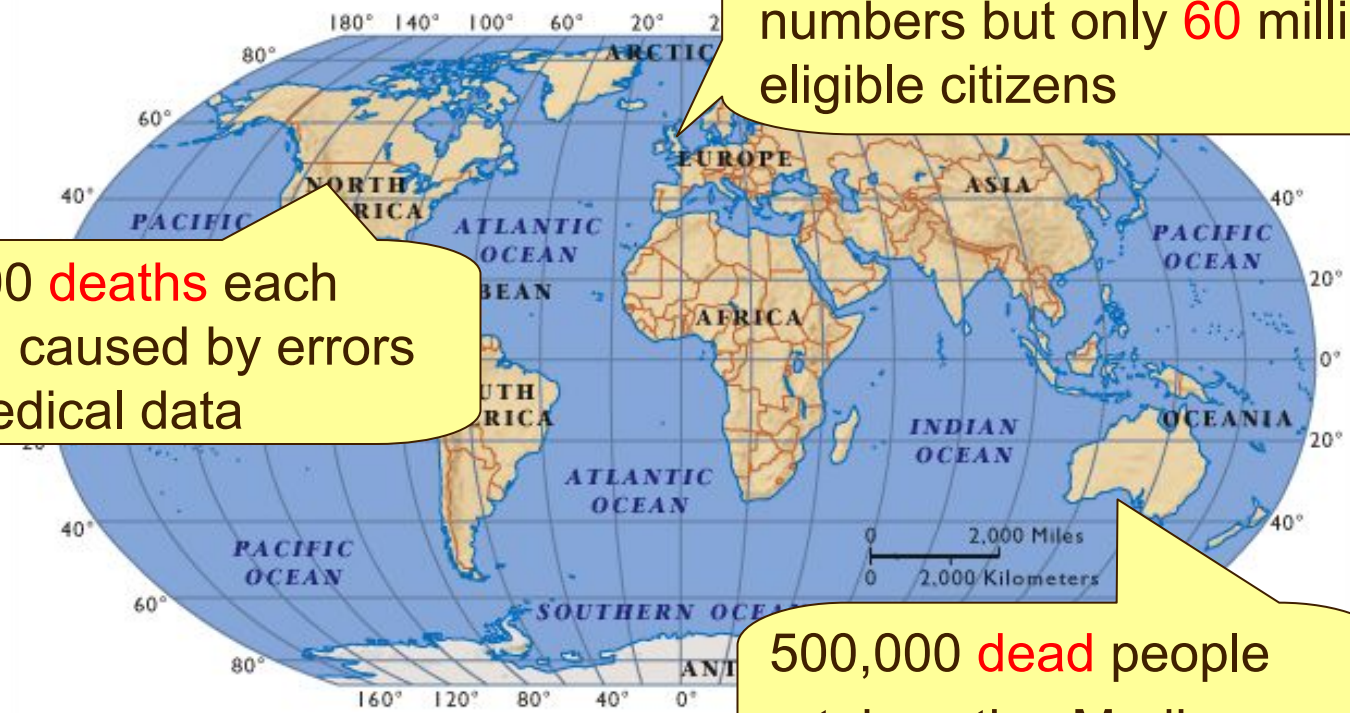
Digital Data is Exploding

According to IBM **90%** of the worlds information…

…was created in the last 2 years

Data in real-life is often dirty

81 million National Insurance numbers but only 60 million eligible citizens



98000 deaths each year, caused by errors in medical data

500,000 dead people retain active Medicare cards

Data error rates in industry: 30%

Dirty data: inconsistent, inaccurate, incomplete, stale

Veracity (quality & trust)

Data = quantity + quality

When we talk about big data, we typically mean its quantity:

- ✓ What capacity of a system provides to cope with the sheer size of the data?
- ✓ Is a query feasible on big data within our available resources?
- ✓ How can we make our queries tractable on big data?
- ✓ . . .

Can we trust the answers to our queries?

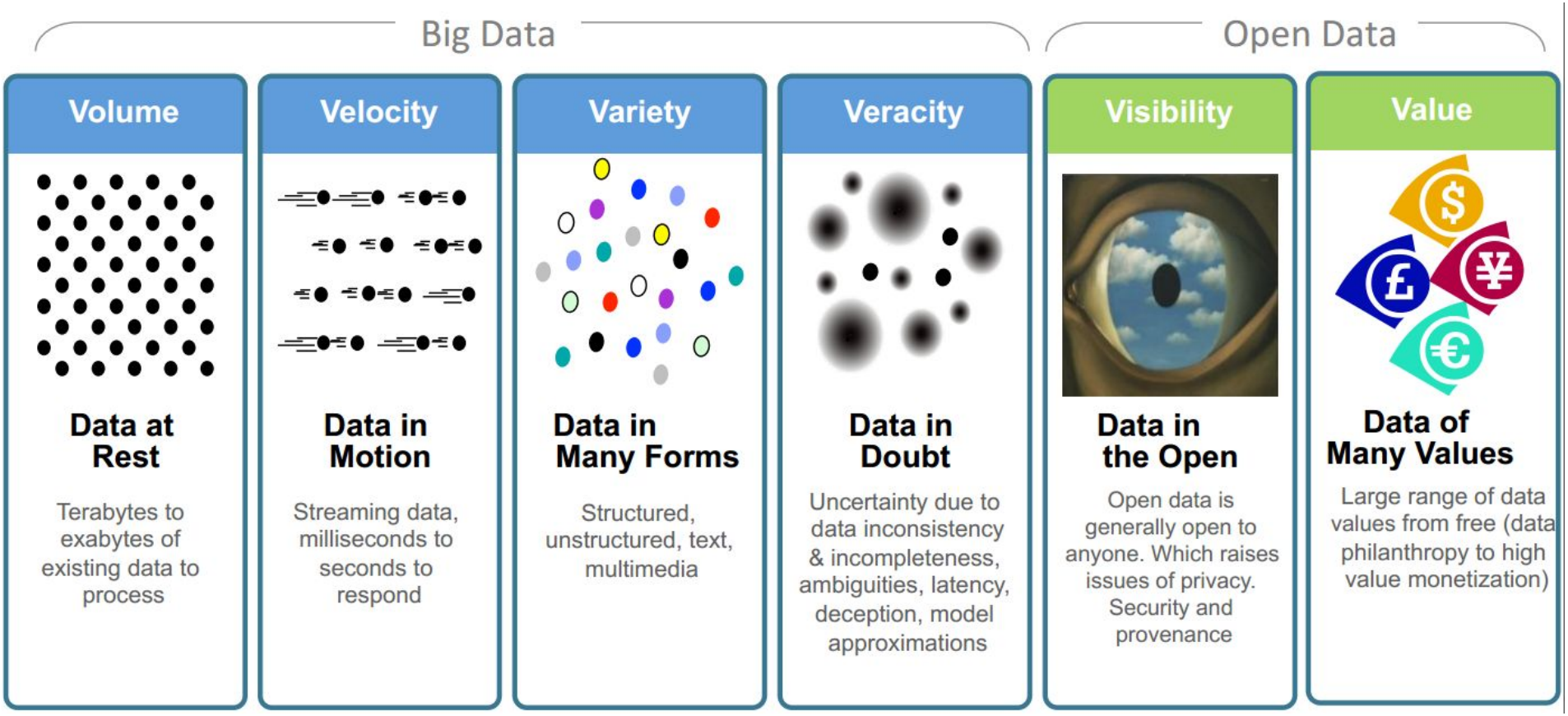
- ✓ Dirty data routinely lead to misleading financial reports, strategic business planning decision ⇒ loss of revenue, credibility and customers, disastrous consequences

The study of data quality is as important as data quantity

Value

- Big data is meaningless if it does not provide value toward some meaningful goal

Big Data: 6V in Summary



Why Study Big Data Technologies?

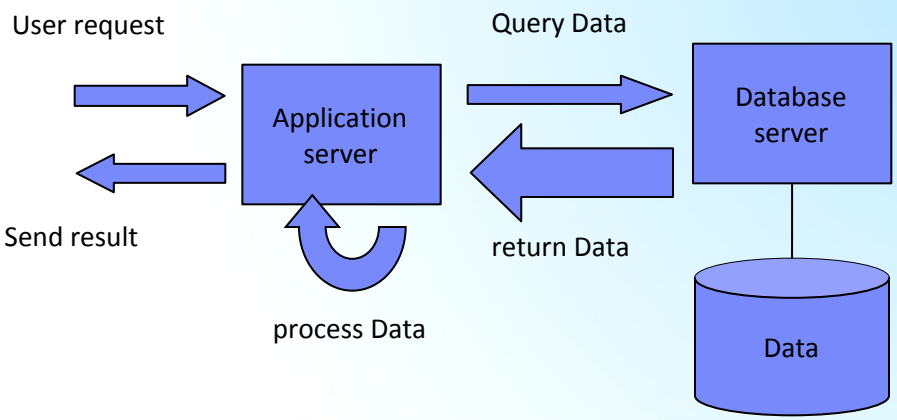
- The hottest topic in both research and industry
- Highly demanded in real world
- A promising future career
  - Research and development of big data systems:
    Distributed systems (eg, Hadoop), visualization tools, data warehouse, OLAP, data integration, data quality control, …
  - Big data applications:
    social marketing, healthcare, …
  - Data analysis: to get values out of big data
    discovering and applying patterns, predicative analysis, business intelligence, privacy and security, …

Demand for Big data skills

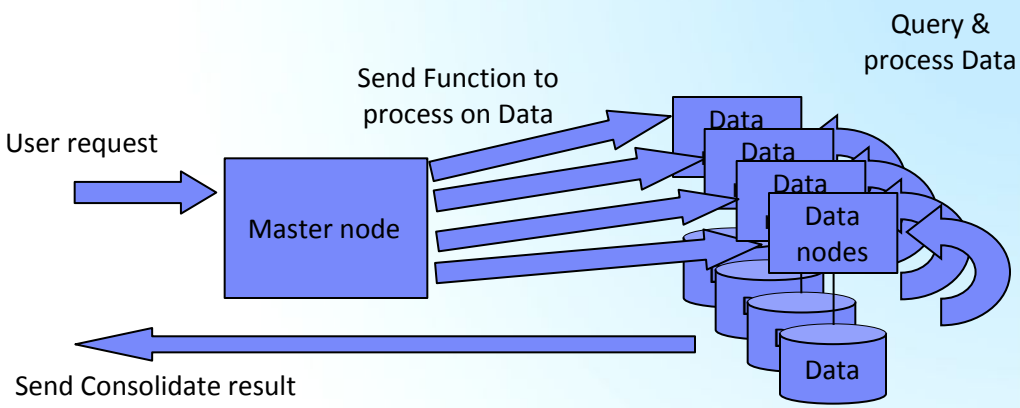
By **2020 16.4 Million** IT jobs will be created to support Big Data – generating **5.9 million** jobs in the United States

Big Data : why is it possible Now ?

I. Traditional approach : **Data to Function**



I. I. Big Data approach : **Function to Data**



I.Traditional approach

I. Application server and Database server are separate
II. Data can be on multiple servers
III. Analysis Program can run on multiple Application servers
IV. Network is still at the middle
V. Data have to go through the network

- Big Data Approach
  - ➢ Analysis Program runs on the data: on Data Node
  - ➢ Only the Analysis Program are have to go through the network
  - ➢ Analysis Program need to be MapReduce aware
  - ➢ Highly Scalable :
    - ➢ 1000s Nodes
    - ➢ Petabytes and more

Big Data Tools

Open Source

Machine Learning

mllib, mahout, Apache SINGA, MADlib, Aerosolve, torch, Caffe, CNTK, TensorFlow, learn, FeatureFu, WEKA, jupyter, VELES, DIMSUM, DL4J

Search

elasticsearch, Solr, Lucene

Security

Apache Ranger

Visualization

Zeppelin

Framework

hadoop HDFS, hadoop MapReduce, YARN, MESOS, Spark, TEZ, Apache Kylin, Flink, CDAP

Data Access

accumulo, APACHE HBASE, mongoDB, cassandra, sqoop, kafka, Flume, SciDB, nifi, CouchDB, riak, OPENTSDB

Query / Data Flow

HIVE, SLAMDATA, APACHE DRILL, Google Cloud Dataflow

Coordination

talend, OOZIE, Apache Zookeeper, Apache Ambari

Real-Time

STORM, Spark, APEX, Flink, TACHYON, druid

Stat Tools

R, Scala, NumPy, SciPy

One popular solution: Hadoop

Hadoop Cluster at Yahoo! (Credit: Yahoo)